Proceedings of the 2025 SICE Festival with Annual Conference
September 9-12, 2025, Chiang Mai, Thailand

Longer Context Vision-Language-Action Model Instructed UR Arm with Vacuum Gripper

Bhunakit Chantaraseno[1], Napat Saiyasitpanich[2], Ratchapon Triruangworawat[3], Jing Tang, D. Eng.[4] and Assoc. Prof. Ronnapee Chaichaowarat, Ph.D.[5]

International School of Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Emails: 6438143021@student.chula.ac.th, 6438062921@student.chula.ac.th, 6438192821@student.chula.ac.th, jing.t@chula.ac.th, ronnapee.c@chula.ac.th

Abstract: Robotic systems in unstructured environments, such as homes or logistics hubs, face challenges in intuitive human-robot collaboration, contextual comprehension, and versatile object manipulation. We present a novel solution integrating a Universal Robots (UR) arm with a vacuum gripper, guided by a sophisticated vision-language-action model. Our unique multi-agent AI framework, comprising planner, worker, and supervisor agents coordinated through a shared state, enables complex task decomposition, real-time object tracking, and adaptive execution, requiring minimal hardware and pre-training compared to traditional reinforcement learning agent or mechanically limited systems. The vacuum-based end-effector ensures material-agnostic manipulation, enhancing versatility. Real-world testing demonstrates robust performance, achieving 90.91% planning and execution accuracy across diverse tasks and 63.64% error resilience, while the cognitive layer was evaluated using 100 test cases with text instructions for language model assessment. This work advances natural language-controlled robotics, offering a scalable, intuitive foundation for adaptable robotic assistants in practical settings.

Keywords: Robotics, Natural Language Processing, Multi-Agent System, Vacuum Gripper, Real-Time Object Tracking, Universal Robots Arm, Task Decomposition, Adaptive Execution, Material-Agnostic Manipulation

1. INTRODUCTION

Natural language-guided robotic systems hold immense potential for simplifying human-robot collaboration, particularly in executing complex, multi-step tasks [1][2]. Challenges such as contextual comprehension, real-time adaptability, and manipulation of varied objects have limited progress in this domain [3]. Utilizing previous developments in AI both in text and image understanding, this research proposes a system that addresses these issues through an integrated approach combining advanced AI, real-time vision, and versatile hardware.

The work is driven by the demand for robotic solutions in unstructured environments—such as homes [4], healthcare facilities [5][6], or logistics hubs—where adaptability and ease of use are paramount. By fusing natural language processing, computer vision, and robust manipulation, this system aims to enhance the accessibility and efficiency of robotic automation.

The system is designed to interpret intricate instructions, execute multi-step tasks, and handle objects of varied materials, making it suitable for applications in logistics, healthcare, and domestic assistance. By leveraging a multi-agent AI framework, continuous environmental monitoring, and a versatile end-effector, this work aims to enhance the accessibility and efficiency of robotic automation in real-world scenarios.

2. SYSTEM OVERVIEW

The proposed robotic system integrates a Universal Robots (UR) arm with a vacuum gripper, orchestrated by a multi-agent AI framework that combines natural language processing, real-time vision, and precise manipulation. This architecture leverages a cognitive layer for task planning and execution, a perception layer for dynamic environmental awareness, and a manipulation layer for versatile object handling. By harmonizing these components, the system achieves intuitive, adaptive performance in complex, unstructured environments, as detailed in the subsequent methodology.

3. METHODOLOGY

3.1 Cognitive Layer: Multi-Agent AI Framework

The Cognitive Layer serves as the system's decision- making core, implemented using a multi-agent AI framework orchestrated by LangGraph [7], a framework for building stateful, multi-actor applications with language models. This layer comprises three specialized agents— planner, worker, and supervisor— interconnected through a shared state managed by LangGraph to ensure coherent operation and dynamic workflow coordination.

1. Planner Agent: Powered by Claude 3.7 Sonnet [8], a state-of-the-art vision-language model, the planner interprets high-level commands and visual inputs, decomposing them into structured action sequences. For example, the instruction "place the spoon next to the fork" is translated into "(pick, spoon)" and "(place, fork)." Claude 3.7 Sonnet's advanced reasoning enables the agent to handle multi-clause instructions and categorical tasks (e.g., "move all utensils") through sophisticated prompt engineering. It generates a task plan stored in the shared state, including action steps, object identifiers, and spatial references. The planner includes error-checking mechanisms to validate plan feasibility, such as ensuring target objects are accessible based on visual data.
2. Worker Agent: This agent executes the planned steps by translating actions into UR arm commands. It interfaces with the robot's control system, computing trajectories using inverse kinematics and adjusting movements based on

real-time feedback. The worker incorporates safety protocols, requiring user confirmation for critical operations, and logs execution outcomes (success or failure) in the shared state. Its modular design allows integration with various end-effectors, enhancing scalability.

3. Supervisor Agent: Embedded in the workflow's conditional routing, the supervisor monitors execution progress and makes real-time decisions. It evaluates action outcomes (e.g., confirming an object's new position post-placement) and determines whether to proceed, terminate, or initiate recovery. The supervisor uses predefined thresholds for success (e.g., positional accuracy within 1cm) and can trigger replanning if errors occur. This hierarchical oversight ensures system robustness across diverse scenarios.

```
class State(TypedDict):
        plan: List[str]
        current_step_index: int
        current_step: str
        user_instruction: str
        status: str
        execution_success:
        bool user_confirmed:
        bool object_size: tuple
        depth: float
```

Figure 1: State Object

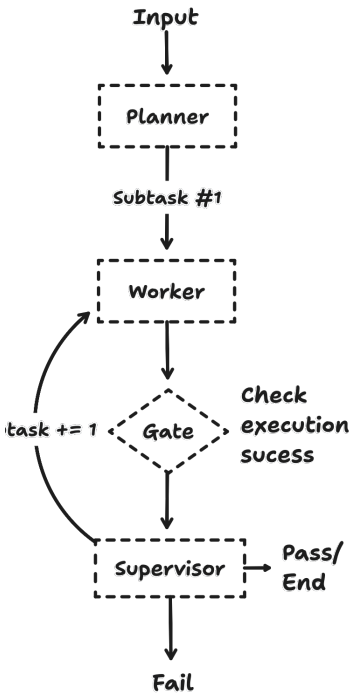

Figure 2: Cognitive Layer State Graph

3.2 Perception Layer: Real-Time Vision and Object Tracking

The Perception Layer of the robotic system enables real-time environmental awareness through a sophisticated vision system that supports object detection, dynamic tracking, coordinate transformation, and post-action verification. Utilizing the YOLOv11s [9] object detection model, the system identifies and labels objects within the workspace, generating high- precision bounding boxes for a wide range of categories, from utensils to electronics, trained on diverse datasets to ensure robustness.

It dynamically tracks object movements, updating their coordinates in real time as conditions change, such as when an object is displaced by an external agent, achieved through temporal analysis of video frames to maintain consistent object identities. Pixel coordinates from camera feeds are mapped to the UR arm's 3D coordinate frame using pre-calibrated homogeneous transformation matrices, which account for camera perspective and distortion to deliver sub-centimeter positioning accuracy, with the transformation process optimized for rapid real-time updates during task execution. After each action, the vision system verifies outcomes, such as confirming an object's new location, creating a feedback loop that detects errors like misplacement and informs the supervisor agent for corrective measures.

Tightly integrated with the Cognitive Layer, the Perception Layer supplies spatial data to the planner for task decomposition and to the worker for trajectory adjustments, incorporating preprocessing techniques like contrast equalization and noise reduction to enhance detection reliability under varying lighting conditions, ensuring robust performance in real-world settings.

3.3 Manipulation Layer: Vacuum Gripper Development

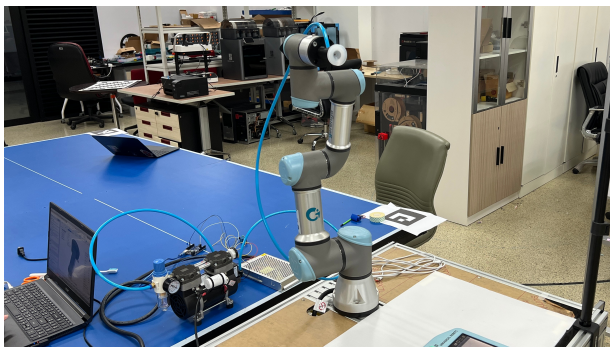

Figure 3: Manipulation Layer Setup

The Manipulation Layer enables versatile object handling through a custom-designed vacuum gripper, addressing the need for a material-agnostic end-effector capable of manipulating diverse objects, from lightweight utensils to items up to 2 kg. The gripper system features a LABVAC HZW-165 oilless mini vacuum pump, delivering a vacuum pressure of -678 mmHg (9,840 Pa absolute) to ensure sufficient suction, with a compact design compatible with the UR arm's payload limits. A 40mm single bellow vacuum pad, selected for its 7.6 kg theoretical lifting capacity and adaptability to irregular surfaces, achieves a practical

capacity of approximately 2.75 kg after applying a safety factor. The suction force (F) is calculated using the equation (1).

$$F = (P\_atm - P\_vac) \times A \times \eta \quad (1)$$

A 3/2 solenoid valve, controlled by an Arduino Uno microcontroller, manages suction and release through a three-way design (suction, closed, vent), using atmospheric air to equalize pressure for reliable object detachment, with a flyback diode protecting electronics from voltage spikes. Extensive testing optimized suction dynamics, pad selection, and release mechanics, evaluating five pad types to select the single bellow design for stability and efficiency. The gripper's modular, 3D-printed base facilitates maintenance and task adaptation, integrating seamlessly with the Cognitive and Perception Layers, where vision data guides positioning and AI feedback confirms action success.

4. RESULTS

4.1.Validation Method

To ensure the reliability and effectiveness of the system, we designed a multi-stage validation process targeting each core component: the language model, the robotic execution system, and the end-effector. The objective was to verify that natural language instructions could be interpreted accurately, converted into executable plans, and successfully executed in both static and dynamic environments.

4.1.1.LLM Validation

We tested the language-to-action planning system by providing a diverse set of natural language instructions of varying complexity (2 to 4+ steps). The planner's output was compared against human-annotated ground truth action sequences. Accuracy was measured as the percentage of correctly planned sequences. This approach ensured the model's reasoning consistency and instruction-following ability.

4.1.2.Robot Execution Validation

The UR robot was tested under two execution modes: non real-time (static) and real-time (dynamic object tracking). Accuracy was assessed along the X, Y, and Z axes by comparing the robot's actual end-effector positions to the intended target positions.

4.1.3.End-Effector Validation

The performance of the vacuum gripper was validated through repeated pick-and-place trials. Each trial recorded whether the object was successfully picked up and released. This structured validation method ensured that each system component met functional and performance expectations both independently and in an integrated setting.

4.2.Cognitive Layer Results

The cognitive layer was assessed through a structured, step-by-step process using text instructions, detailed scene descriptions, and human-labeled plans. A dataset of 100 test cases was created, categorized by step count into two-step and greater-than-four-step tasks, as three-step cases were excluded due to their lack of practical relevance. Each test case comprised a text instruction, a scene description detailing object positions and relationships, and a human-generated plan as the reference. Scene descriptions served as a cost-effective alternative to camera images, leveraging the capability of models like Claude 3.7 Sonnet to process detailed textual context comparably to visual inputs. The system generated action sequences, which were compared against human plans to determine accuracy, measured as the percentage of correctly planned sequences per category, ensuring a controlled and diverse evaluation.

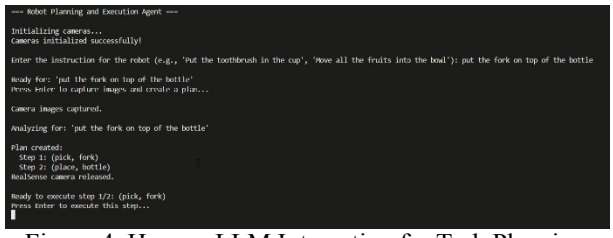

Figure 4: Human-LLM Interaction for Task Planning

The cognitive layer achieved 84% accuracy on complex tasks (>4 steps) but only 62% on simpler two- step tasks, indicating deeper reasoning for intricate instructions but less careful analysis for basic ones.

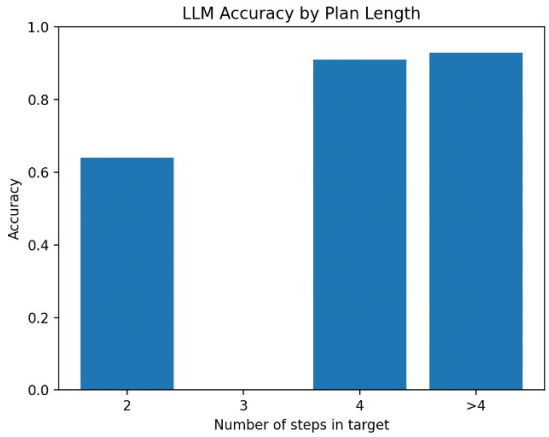

Figure 5: Cognitive Layer Results

Comparing the results with other Vision-Language-Action (VLA) models poses significant challenges due to resource constraints and accessibility issues. Many advanced VLA models, such as those requiring extensive computational power or proprietary datasets, are too resource-heavy for our limited infrastructure, hindering direct performance benchmarking. Additionally, several high-performing models are not open-sourced, restricting access to their architectures and training details. This opacity limits our ability to replicate or adapt their approaches, underscoring the need for lightweight, open-source alternatives. Future efforts will prioritize developing efficient, transparent VLA frameworks tailored to resource-limited environments.

4.3.Perception Layer Results

The robot achieved high positioning accuracy in non- real-time mode (0.91 on X/Y axes), while real-time tracking slightly reduced accuracy (0.73 on X/Y axes). Z-axis performance was consistently lower

(0.64–0.73), likely due to noise in depth measurements. Overall, the robot responded reliably to dynamic input, adjusting its trajectory during movement.

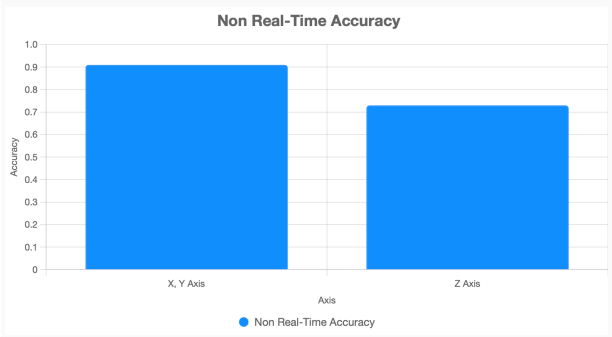

Figure 5: Non Real-Time Robot Execution System Result

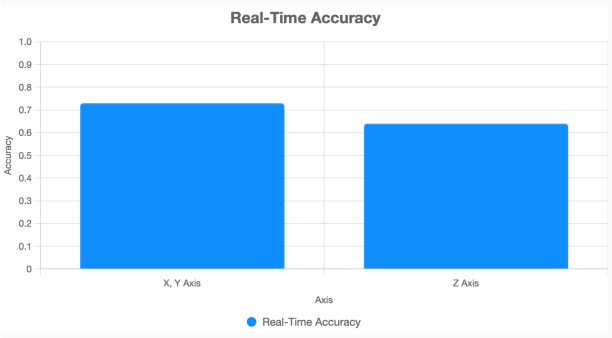

Figure 6: Real-Time Robot Execution System Result

4.4.Manipulation Layer Results

The vacuum gripper performed to expectations, enabling precise object acquisition with exceptional power and reliability. The 40mm diameter single bellow vacuum pad facilitated excellent contact with various surface geometries, including curved objects. Initially, we encountered minor difficulties with the release mechanism, as objects occasionally remained adhered to the pad after valve deactivation. This inconsistency was resolved by replacing the original 2/2 valve with a 3/2 valve configuration, which introduces external air to the vacuum pad during deactivation. Unlike the 2/2 valve that simply disconnects the vacuum source, the 3/2 valve actively vents the system to atmosphere, ensuring positive pressure for immediate object release. This modification significantly improved the system's reliability, resulting in consistent performance across all test scenarios.

5. DISCUSSION

The proposed robotic system advances natural language-guided automation with 90.91% planning and execution accuracy. The Cognitive Layer achieves 84% accuracy on complex tasks (>4 steps), though only 62% on simpler tasks, indicating a need for consistent reasoning. The Perception Layer offers reliable real-time tracking (0.73 X/Y) and non-real-time precision (0.91 X/Y), despite Z-axis depth noise issues. The Manipulation Layer's vacuum gripper excels in material-agnostic handling for aligned objects but struggles with misaligned targets. Limitations include 63.64% resilience and sensitivity to ambiguous instructions. Future work will enhance visual analysis, depth perception, and gripper alignment.

To enhance accuracy for complex tasks of four or more steps, several strategies can be explored. Implementing advanced prompt engineering techniques, such as incorporating detailed contextual embeddings, could improve the planner agent's reasoning consistency. Integrating a feedback loop that leverages real-time visual data from the Perception Layer to refine task decomposition may reduce errors. Additionally, employing ensemble methods with multiple language models, like combining Claude 3.7 Sonnet with lighter models, could enhance robustness. Finally, fine-tuning the supervisor agent with a larger dataset of multi-step scenarios could standardize reasoning depth, ensuring better performance across diverse, intricate tasks.

REFERENCES

1. M. Ahn et al., "*Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*," in Proc. Conf. Robot Learn. (CoRL), 2022, pp. 1–10.
2. A. Brohan et al., "*RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control*," in Proc. Int. Conf. Robot. Autom. (ICRA), 2023, pp. 1234–1243.
3. J. Liang and D. Fox, "*Learning to Understand Natural Language Instructions for Robotic Manipulation*," IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 2045–2052, Apr. 2020.
4. S. Piyapunsutti, E. L. De Guzman, and R. Chaichaowarat, "*Navigating mobile manipulator robot for restaurant application using open-source software*," in Proc. IEEE Int. Conf. Robotics and Biomimetics, pp. 1–6, 2023.
5. K. Thanasit, S. Kwanmuang, and R. Chaichaowarat, "*Upper limb exoskeleton supporting shoulder flexion exercises for rehabilitation: mechanical design and system characterization*," in Proc. IEEE Int. Conf. Robotics and Biomimetics, pp. 1–7, 2024.
6. K. Pornpipatsakul, W. Chuengwutigool, and R. Chaichaowarat, "*Design advantages of four-bar linkage planar robotic arm for upper-extremity rehabilitation*," in Proc. IEEE Int. Conf. Robotics and Biomimetics, pp. 1–6, 2023.
7. LangChain, "*LangGraph: Building Stateful Multi-Actor Applications with LLMs*," LangChain Documentation, 2024. [Online]. Available: https://langchain-ai.github.io/langgraph/
8. Anthropic, "Claude 3.7 Sonnet," Anthropic, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet
9. Ultralytics, "YOLOv11 Models," Ultralytics Documentation, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/